Improving Pattern Recognition Based Pharmacological Drug Selection Through ROC Analysis*

W. Díaz[1], María José Castro[2], F.J. Ferri[1,**], F. Pérez[3], and M. Murcia[3]

[1] Dept. Informàtica, Universitat de València, 46100 Burjassot, Spain
ferri@uv.es
[2] Dept. Sistemes Informàtics i Computació, Universitat Politècnica de València 46071 València, Spain
[3] Dept. Química Física, F. Farmàcia, Universitat de València, 46100 Burjassot, Spain

Abstract. The design of new medical drugs is a very complex process in which combinatorial chemistry techniques are used. The goal consists of discriminating between molecular compounds exhibiting or not certain pharmacological activities. Different machine learning approaches have been recently applied to different drug design problems leading to competitive results in pointing at particular compounds with high probability of exhibiting activity. The present work first deeps into the natural trade-off between accuracy in the much less populated active group and false alarm rate which could lead to too many expensive laboratory tests. Preliminary results show how different classification techniques are suited for this particular problem and throw light to keep improving the results by considering also the acceptance/rejection trade-off.

1 Introduction

The design of new medical drugs with desired chemical properties is a challenging and very important problem in the pharmaceutical industry. The traditional approach for formulating new compounds requires the designer to test a very large number of molecular compounds, to select them in a blind way, and to look for the desired pharmacological property. Therefore, it is very useful to have tools to discriminate the pharmacological activity of a given molecular compound so that the laboratory experiments can be directed to those molecular groups in which there is a high probability of finding new compounds with the desired properties.

All methods developed for this purpose are based on the fact that the activity of a molecule derives from its structure and therefore it is possible to find a relationship between this structure and the properties that the molecule exhibits [9]. Thus, the way the molecular structure is represented has special relevance.

* This work has been partially funded by spanish project TIC2003-08496.
** Contacting author

A. Sanfeliu et al. (Eds.): CIARP 2004, LNCS 3287, pp. 621–628, 2004.

In Chemical Graph Theory, molecular structures are represented as doubly labelled graphs which can be conveniently characterized by a number of specific topological indices [6]. In this work, a reduced set of 62 topological indices [7] are considered.

These or similar representations have already been applied to different discrimination problems in drug design (analgesic, antidiabetic, antibacterial, etc.) but the cost/benefit problem and the corresponding discrimination thresholds have always been selected at hand based on previous a priori knowledge on the particular problem.

This paper describes several Machine Learning approaches (LDA analysis, naive Bayes classifier and neural networks) to solve a particular problem of property discrimination (antibacterial activity) based on the structural representation of the molecule. A detailed analysis is then performed in order to determine the suitability and adaptability of these methods for the particular task from the point of view of the cost/benefit trade-off.

2 The Molecular Representation Space

As an alternative to the methods based on the "exact" description of the electronic properties of a molecule calculated by mechanical-quantum methods, the molecular topology describes the molecule as a set of indices. These topological indices are numerical descriptors that encode information about the number of atoms and their structural environment. This representation is derived from the hydrogen-suppressed molecular formula seen as a graph and it requires a relatively low calculation effort [1, 2, 6].

The molecular topology considers a molecule as a planar graph where atoms are represented by vertices and chemical bonds are represented by edges. The chosen set of molecular descriptors should adequately capture the phenomena underlying the properties of the compound. In this and other related works, a set of 62 indices has been selected [7, 8, 10]. Fourteen of these indices are related to the molecular attributes of the compound; for example, the total number of atoms of a certain element (carbon, nitrogen, oxygen, sulphur, fluorine, chlorine, ...), the total number of bonds of a certain type (simple, double or triple), the number of atoms with a specific vertex degree, distance between the bonds, etc...

The remaining forty-eight indices include different topological information, such as the number of double bonds at distance 1 or 2, and the minimum distance between pairs of atoms, which are counted as the number of bonds between atoms. These indices are classified into six groups which are associated to the most frequent elements that constitute the molecules with pharmacological activity: nitrogen, oxygen, sulphur, fluorine, chlorine, bromine, and a general group in which the distances between pairs of atoms are considered without identifying the type of atom.

This molecular representation has shown its ability for discriminating and predicting different kinds of pharmacological properties. Nevertheless, it is known

that certain indices are more important that others for detecting particular cases. For example, it has been shown that only eight out of the above topological indices are enough to predict antibacterial activity with about 80% accuracy (and about 90% inactivity accuracy or 10% false alarm rate) [8].

As we are trying to extract conclusions as general as possible from the empirical evaluation carried out, we will not use this kind of information in the experiments and the whole set of topological descriptors will be considered. This means that the methods with a natural tendency to extract a few good features (as multilayer neural networks) will have more chances to obtain better results.

3 The Antibacterial Activity Discrimination Problem

The particular discrimination problem was to determine whether a molecule has antibacterial activity or not. To this end, three different classification thechniques have been considered: LDA analysis [8], Multilayer Perceptrons [3] and Gaussian Naive Bayes [4] as a reference.

A dataset of 434 samples with potential pharmacological activity has been used in this work. This particular dataset is basically balanced which is not representative of the a priori probability of antibacterial activity in real pharmacological design trials.

In order to obtain results as significant as possible, repeated cross-validation has been used to compute all accuracies. In particular, the dataset was split five times into training and test set in a 70%-30% proportion and the corresponding results averaged. Every partition has been performed randomly, taking into account that the percentages of active and inactive samples were the same as in the original sample.

Instead of performing new experiments, the results with LDA analysis have been directly taken from a previous work [8] which used a different (and slightly more representative) dataset and can be considered as the best results to date for this particular problem.

New experiments have been performed with a naive Bayes classifier [4] to use them as a reference. Finally, multilayer perceptrons (MLPs) were used to discriminate antibacterial properties of the molecules.

The 62 topological indices were used to obtain feature vectors in which values were linearly normalized to the interval $[0, 1]$ in an independent way. As in previous works [8], each feature vector was labeled with 1 (indicating that the molecule has antibacterial properties) and -1 (the molecule is inactive).

The training of the MLPs were carried out using the neural software package "SNNS: Stuttgart Neural Network Simulator" [12]. The network topology, training algorithm and parameter settings were chosen from a previous work [3] which was not particularly aimed at looking for antibacterial activity.

More specifically, the results presented have been obtained by using the standard Backpropagation algorithm with a learning rate equal to 0.05 and a topology of 62 input units, 2 hidden units and one output unit. The hyperbolic tangent function was used in order to keep outputs in the interval $[-1, 1]$ as in the LDA analysis [8].

The Naive Bayes implementation used was taken from the data mining and machine learning package Weka [11]. The outputs were normalized also to the $[-1, 1]$ interval to keep all results and discrimination thresholds comparable.

4 The Cost Benefit Trade-Off

It is important to note that we are interested not only in achieving a high accuracy in classification but also a convenient compromise between true positive and false alarm rates. The high economical costs due to the pharmacological tests on each candidate molecule in drugs research makes an important issue to keep the number of false positives as low as possible, even if this implies to reject some true positives.

Table 1. Confusion matrix corresponding to LDA [8]. All undetermineds are taken from molecules predicted as active. Consequently, accuracies for both active and inactive groups without a reject option are 81.95% and 92.58%, respectively.

	predicted		
	active	undetermined	inactive
active	70.83%	11.12%	18.05%
inactive	5.94%	1.48%	92.58%

As already mentioned, orientative figures for previous results on this problem are about 80% of true positives and 10% of false positives. The whole confusion matrix obtained for antibacterial prediction using LDA in [8] is shown in Table 1. Nevertheless, the emphasis of the experimentation in this work is on studying how the classifiers behave as we constrain one of the two above figures. In other words, we are interested in obtaining a Receiver Operating Characteristic (ROC) curve and looking for the best parameter settings of each classification scheme.

Given a particular classifier whose output consists of a continuous value in a specified interval (as in the cases considered in this work), the ROC curve is defined as the plot of the true positive rate (TP) against false positive rate (FP) considering the threshold used in the classifier as a parameter. The so-called ROC space is given by all possible results of such a classifier in the form (FP,TP). The performance of any classifier (with the corresponding threshold included) can be represented by a point in the ROC space. ROC curves move from the "all-inactive" point (0,0) which corresponds to the highest value of the threshold to the "all-active" point (1,1) given by the lowest value for the threshold. The straight line between these two trivial points in the ROC space corresponds to the family of random classifiers with different a priori probabilities for each class. The more a ROC curve separates from this line, the better the corresponding classification scheme is. As ROC curves move away from this line, they approach the best possible particular result that corresponds to the point (0,1) in the ROC space which means no false alarms and highest possible accuracy in the active class.

The ROC curve is a perfect tool to find the best trade-off between true positives and false positives and to compare classifiers in a range of different situations. A number techniques to obtain different measures from ROC curves have been also developed.

5 Experiments, Results and Discussion

The above mentioned classification methods have been applied to the training sets taken from the available data set and the corresponding (continuous) outputs have been obtained for the test data. For each partition into train and test, a ROC curve is obtained. Figure 1 shows the corresponding averaged ROC curves for the three classification schemes considered. In the particular case of LDA, the curve corresponds to a unique partition in train and test data as explained in [8].

The ROC curve corresponding to LDA shows that, apart from the already mentioned $(0.08, 0.82)$ classifier with discrimination threshold set to zero, there are other possible convenient classifiers as the one with point $(0.005, 0.72)$ in ROC space which uses a higher discrimination threshold, namely 0.61.

More importantly, the (particularly simple) Multilayer Perceptron considered, significantly outperforms the LDA results along the whole range of the curve. Particular (averaged) results that can be mentioned are $(0.05, 0.95)$ and $(0.10, 0.82)$ that involve discrimination thresholds very close to 1.

Finally, results obtained with the Gaussian Naive Bayes are clearly and significantly worse than the others and have been included as a reference only.

The threshold averaged ROC curves have been computed as explained in [5]. The corresponding intervals for a 95% confidence level have also been computed and are shown in Figure 2 for the case of Multilayer Perceptron.

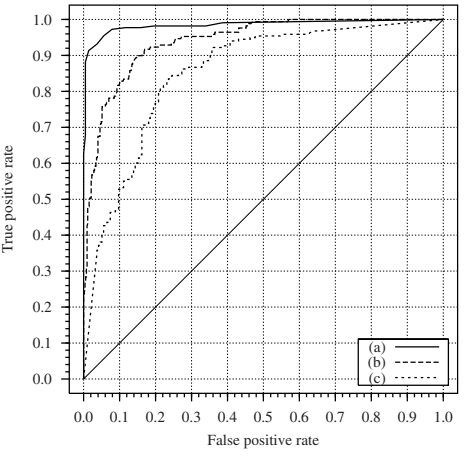

Fig. 1. Averaged ROC curves for a) Multilayer Perceptron, b) LDA, and c) Gaussian Naive Bayes.

Further experimentation including a reject option as in [8] shows that there is still room for improving these results. In particular, the results in Table 1 are obtained by rejecting outputs in the interval $[-0.5, 0.5]$ which means rejecting about 7% of the cases if the data set was balanced. We have carried out an exhaustive computation of ROC curves for all possible percentages of rejection for a particular dataset partition in the case of Multilayer Perceptron. From all curves (in fact, classifiers) obtained, we have selected those with 0% (the original ROC curve), 15% and 20% rejection rate. Results for two particular partitions of the data are shown in Figure 3. As can be seen, the curves are getting better as rejection rate increases. Although these results are still preliminary and valid only for a specific partition of the (by the way quite reduced) data set, this family of curves could eventually be used to look for more convenient classifiers for the problem at hand.

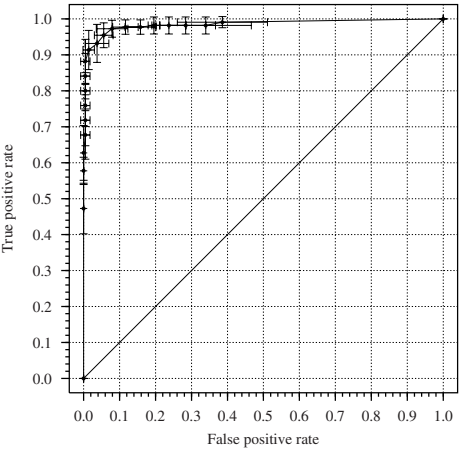

Fig. 2. Averaged ROC curve corresponding to the multilayer perceptron approach along with 95% confidence intervals.

6 Concluding Remarks and Further Work

In this work a classical ROC analysis has been performed on a particular drug activity discrimination problem. Preliminary results show that this kind of analysis is very interesting and can significantly improve the overall costs in the whole drug design methodology. In particular, very simple Multilayer Perceptron is shown to significantly improve previously used approaches in a wide range of situations.

In order to obtain more confident results significant also from a pharmacological point of view, the whole experimentation in this work needs to be repeated with a larger and more representative data set. Also, ROC analysis including the whole range of possible rejects is under consideration. This will lead to a

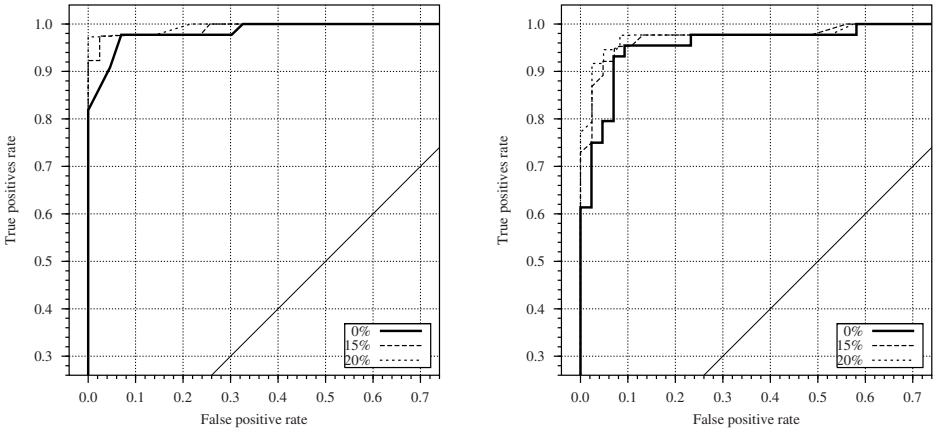

Fig. 3. ROC curves obtained at 0%, 15% and 20% rejection rates for two particular partitions of the available data set using Multilayer Perceptron.

ROC surface into a three dimensional ROC space given by true positive rate, false alarm rate and reject rate and would lead to a full characterization of the discrimination problem in drug design applications.

References

1. A.T. Balaban, I. Motoc, D. Bonchev, and O. Makenyan. Topological indices for structure-activity correlations. *Top. Curr. Chem.*, 114:21–55, 1983.
2. S.C. Basak, S. Bertelsen, and G. Grunwald. Application of graph theoretical parameters in quantifying molecular similarity and structure-activty studies. *J. Chem. Inf. Comput. Sci.*, 34:270–276, 1994.
3. M. J. Castro, W. Díaz, P. Aibar, and J. L. Domínguez. Prediction and Discrimination of Pharmacological Activity by Using Artificial Neural Networks. In F. J. Perales, A. J. C. Campilho, N. Pérez de-la Blanca, and A. Sanfeliu, editors, *Pattern Recognition and Image Analysis*, volume 2652 of *LNCS*, pages 184–192. Springer-Verlag, 2003. ISSN 0302-9743.
4. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley and Sons, second edition, 2001.
5. Tom Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine Learning*, submitted(`http://www.hpl.hp.com/personal/Tom_Fawcett/papers/ROC101.pdf`), 2004.
6. J. Gálvez, R. García-Domenech, J.V. de Julián-Ortiz, and R. Soler. Topological approach to drug design. *J. Chem. Inf. Comput. Sci.*, 35:272–284, 1995.
7. J. Jaén-Oltra, M.T. Salabert-Salvador, F.J. García-March, F. Péz-Giménez, and F. Tomás-Vert. Artificial neural network applied to prediction of fluorquinolone antibacterial activity by topological methods. *J. Med. Chem.*, 43:1143–1148, 2000.

8. M. Murcia-Soler, F. Pérez-Giménez, F.J. García-March, M.T. Salabert-Salvador, W. Díaz-Villanueva, and P. Medina-Casamayor. Discrimination and selection of new potential antibacterial compounds using simple topological descriptors. *J. Mol. Graph. Model.*, 21:375–390, 2003.
9. P.G. Seybold, M. May, and U.A. Bagal. Molecular structure-propertiy relationships. *J. Chem. Educ.*, 64:575–581, 1987.
10. F. Tomás-Vert, F. Pérez-Giménez, M.T. Salabert-Salvador, F.J. García-March, and J. Jaén-Oltra. Artificial neural network applied to the discrimination of antibacterial activity by topological methods. *J. Mol. Struct. (THEOCHEM)*, 504:272–276, 2000.
11. Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2000.
12. A. Zell et al. *SNNS: Stuttgart Neural Network Simulator. User Manual, Version 4.2*. Institute for Parallel and Distributed High Performance Systems, University of Stuttgart, Germany, 1998.